Epidemics 27 (2019) 123–131

Contents lists available at ScienceDirect

Epidemics

journal homepage: www.elsevier.com/locate/epidemics

Analyzing and forecasting the Ebola incidence in North Kivu, the Democratic Republic of the Congo from 2018–19 in real time

Andrei R. Akhmetzhanov[a,1], Hyojung Lee[a,1], Sung-mok Jung[a], Taishi Kayano[a], Baoyin Yuan[a], Hiroshi Nishiura[a,b,*]

[a] Graduate School of Medicine, Hokkaido University, Sapporo, Hokkaido, Japan
[b] CREST, JapanScience and Technology Agency, Saitama, Japan

ARTICLE INFO

Keywords:
Reporting delay
Nowcasting
Forecasting
Prediction
Effective reproduction number

ABSTRACT

During an Ebola virus disease (EVD) outbreak, the analysis and forecasting of the incidence in real time is challenged by reporting of cases, especially the reporting delay. It should be remembered that the latest count of cases is likely underestimated in real time, and moreover, the effective reproduction number, i.e. the average number of secondary cases generated by a single primary case at a given point in time, is also underestimated without proper adjustment. The present study aimed to adjust the reporting delay to appropriately estimate the latest incidence and obtain short-term forecasts from weekly reporting data of EVD in North Kivu, the Democratic Republic of the Congo (DRC). A semi-structured modeling approach was taken, accounting for reporting delay which can depend on time. The mean reporting delay was estimated at 11.6 days (95% CI: 11.3, 11.9) and the standard deviation was estimated to have changed from 26 November 2019 from 8.5–6.0 days. Nowcasting was successfully implemented by account for the time-dependent reporting delay: it mostly contained future observed values within the 95% confidence intervals, but there were failures when the reported incidence abruptly changed over time. Forecasting was also exercised in a similar manner to the nowcasting, while we imposed an extrapolation approach to the effective reproduction number for two future weeks. Moving average of the reproduction numbers for a few weeks prior the latest time of observation outperformed other extrapolations. The information that we can gain from real time (i.e. sequential) update of "situation report" can be considerably improved by integrating the proposed nowcasting and forecasting to the surveillance system.

1. Introduction

Since the first outbreak of Ebola in Sudan, 1976 (Report of a WHO/International Study Team, 1978), there have been recurrent introductions of Ebola virus disease (EVD) from animals to humans. Even though the largest ever epidemic in West Africa from 2014-16 was successfully contained and lessons were learnt, there are still obstacles in managing the EVD outbreaks swiftly. The Democratic Republic of the Congo (DRC), situated in central region of Africa, preserves a natural habitant of bat species and wildlife primate species are also seen in the forest, allowing frequent emergence of EVD in humans via cross-border transmission at the human-animal interface (Barry et al., 2018; Ponce et al., 2019). As a result, the country has experienced two different outbreaks within the same year, 2018. The first outbreak in Équateur province was successfully controlled by late May 2018. The second outbreak in North Kivu and Ituri and the country's tenth Ebola outbreak erupted in July-August and has not been seen to be stopped after nine months. The toll counts reached more than 1200 cases and the outbreak is ranked as the second largest of recorded EVD epidemics in humans across the world as of the end of the year 2018. While the earlier outbreak in DRC, 2018 came to the end, and the Ministry of Health (MOH) and the World Health Organization (WHO) declared the end on 24 July 2018 (WHO, 2018a), and another emergence with a confirmed Ebola case was ironically reported in the same country on 1 August 2018 (WHO, 2018b). The latest outbreak is 10th recorded emergence in the DRC, and due to sustained chains of transmission over an extended period longer than 6 months in east forest zones, there has been a concern toward the cross-border spread to Uganda and Rwanda. WHO revised its risk assessment and elevated the risk at national and regional levels from high to very high on 28 September 2018 (WHO, 2018c).

Due to a difficulty in approaching the regional community, it has been challenging to control the outbreak in rural regions of North Kivu

* Corresponding author at: Graduate School of Medicine, Hokkaido University, Kita 15 Jo Nishi 7 Chome, Kitaku, Sapporo, Japan.
E-mail address: nishiurah@med.hokudai.ac.jp (H. Nishiura).
[1] Equal contribution.

https://doi.org/10.1016/j.epidem.2019.05.002
Received 14 February 2019; Received in revised form 24 April 2019; Accepted 2 May 2019
Available online 03 May 2019

and Ituri from 2018-19. Geographically, different armed groups with conflict have been seen in North Kivu, preventing international organizations from intervening and tracing contacts. On 16 November 2018, once the United Nations Organization Stabilization Mission in the DRC in Beni experienced a violent interference of their contact tracing and vaccination practice, their practice had to be suspended (WHO, 2018d). In the end of 2018, protesters toward the presidential election in late December physically intervened activities of Ebola treatment center in Beni (WHO, 2018e). Consistent diagnosis and reporting as well as contact tracing have been challenged by political instability.

In such a circumstance, the analysis and forecasting of the epidemic in real time is challenged by reporting, especially the reporting delay, as echoed by a recent nonparametric estimation study (Tariq et al., 2019). While the reporting delay has been already recognized as critical in analyzing other epidemic datasets (Reijn et al., 2011; WHO, 2018f), it is even more so in the setting of the latest EVD outbreak in the DRC. The reported latest incidence by the week of illness onset is underestimated due to reporting delay, and therefore, it must be remembered that the latest count of cases is likely underestimated (Hahné et al., 2009; White et al., 2009), and moreover, the effective reproduction number, i.e. the average number of secondary cases generated by a single primary case at a given point in time, is considerably underestimated without proper adjustment (Cowling et al., 2010). Instead, appropriately accounting for the reporting delay, we could offer an avenue to obtain a short-term forecast of the epidemic. Nowcasting and short-term forecasting at a local setting are expected to improve the situation awareness among healthcare experts as well as public health policymakers and local residents (Finger et al., 2018; Funk et al., 2018).

We thus believe that devising a simplistic model to allow a real-time interpretation of the epidemic is of utmost importance. The present study aims to adjust for reporting delay to appropriately estimate the latest incidence and obtain short-term forecasts from coarsely reported incidence data of EVD in the DRC.

2. Materials and methods

2.1. Epidemiological data

The numbers of cases by week of illness onset and also by week of report were extracted from governmental organizations that publicly announced the information, i.e., WHO AFRO External Situation Reports on Ebola virus disease in DRC (WHO AFRO SitRep) (WHO, 2019), WHO Disease outbreak news (WHO DON) (WHO, 2019), and WHO AFRO Weekly bulletins on outbreaks and emergences (WHO AFRO OEW) (WHO AFRO, 2019). The datasets from all three sources rested on the daily update of EVD from the MOH, DRC and were used for estimating the reporting delay. Despite irregularity in the reporting intervals, multiple snapshots of the epidemic curve by illness onset and reporting week for 2–3 times were obtained every week, and including the data up to the first week of April 2019, there have been a total of 82 snapshots obtained. Fig. 1 shows six different snapshots, approximately separated by one month.

Forecast evaluation was performed by comparing the nowcasted/forecasted weekly counts against reported (observed) counts of laboratory confirmed cases by MOH. The forecasting exercise was performed during the first six months of the outbreak. We focused only on the counts of laboratory confirmed cases, because we believe that the process of addition and removal of probable cases cannot be tackled by our modeling approach in a straight forward manner.

We observed that counts of laboratory confirmed cases were different even on the same observation date, greatly varying by snapshot of the epidemic curve (i.e. different by reporting date). Confirmed cases reported in the early phase of the outbreak have been more likely to have involved a discrepancy in their counts among different snapshots, especially between epidemic curves before and after mid- to late-November. There has been a correction in temporally distributing cases according to their time of illness onset through reconciliation of records and also by removal of duplicated cases (WHO, 2019). Such correction took place in other calendar times too, but the abruptness was not as large as that seen in November.

2.2. Modelling delay distribution

Firstly, we estimate the reporting delay distribution by using an equation for forward calculation of the epidemic snapshot (Tsuzuki et al., 2017). To project the number of unreported cases at a given latest time point, we first estimated the distribution of the reporting delay, exploring a total of 82 snapshots of epidemic curve over time comparing them with later curves (Tsuzuki et al., 2017). Supposing that x_k, $k = 1,2, ..., m$, represent the reporting time, and let us denote the latest reporting time x', all datapoints in the curve x are given by weekly counts $c_{t,x}$ that are arranged by the week of illness onset. The expected number of reported cases in week t can be written as follows:

$$E(c_{t,x}; \theta) = j_t F(\tau; \theta), \tag{1}$$

where $E(.)$ represents the expected value, $\tau = x - t + 1$ (day) is the gap time from illness onset data to the reporting date x (with one more day due to possible report on the latest day), and $F(\tau; \theta)$ represents the cumulative distribution function of the delay from illness onset to reporting with a set of parameters θ that we would like to estimate. j_t is the actual number of cases by illness onset time t. We examined four possible statistical distributions as candidates for describing the delay function F, i.e., Weibull, gamma, log-normal, and exponential distributions, and chose the one with the smallest Akaike Information Criterion with a correction for small sample size (AICc) value.

For each epidemic curve x, we reassert the expectation $E(c_{t,x'}; \theta)$ as:

$$E(c_{t,x'}; \theta) = c_{t,x'} F(x' - t + 1; \theta)/F(x - t + 1; \theta), \tag{2}$$

to avoid an explicit identification of j_t, and estimate parameters θ by maximizing the likelihood that assumes that the latest number of cases by week of illness onset follows a Poisson distribution:

$$L(\theta; c_{t,x}) = \prod_{k=1}^{m} \prod_{t=t_0}^{x(k)} \frac{\left(c_{t,x(k)} \frac{F(x' - t + 1; \theta)}{F(x(k) - t + 1; \theta)}\right)^{c_{t,x'}} \exp\left(-c_{t,x(k)} \frac{F(x' - t + 1; \theta)}{F(x(k) - t + 1; \theta)}\right)}{c_{t,x'}!}, \tag{3}$$

where $x(k)$ represents the latest date of reporting for an epidemic curve k and t_0 is the first observed date of reporting for all epidemic curves (i.e., t_0 = 30 April 2018) and m is the number of epidemic curves (m = 81). Depending on whether the delay distribution is stable over the course of the outbreak, or varies with time, we examined the distribution with or without a change point in statistical moment of the distribution. If the change occurs in a week, the reporting delay of cases with the illness onset before the change point follows a distribution with parameters $\theta = \theta_1$, and all cases who developed illness after the change point are reported with the delay that is governed by parameters $\theta = \theta_2 \neq \theta_1$. For a given x', we compare two models with a stable delay distribution and the one that accounts for one change point, again comparing the AIC_C.

The maximized likelihood method was employed to obtain point estimates $\theta = \hat{\theta}$ as well as the Hessian matrix $H(\hat{\theta})$. The latter allowed us to construct confidence intervals (CI) by implementing parametric bootstrapping. To do so, we designed a dataset of parameters θ sampled from the multivariate normal distribution with the mean $\hat{\theta}$, and took 2.5th and 97.5th percentile points of the simulated distribution as the 95% CI.

2.3. Nowcasting implementation

We assessed the number of unreported cases by sampling them from

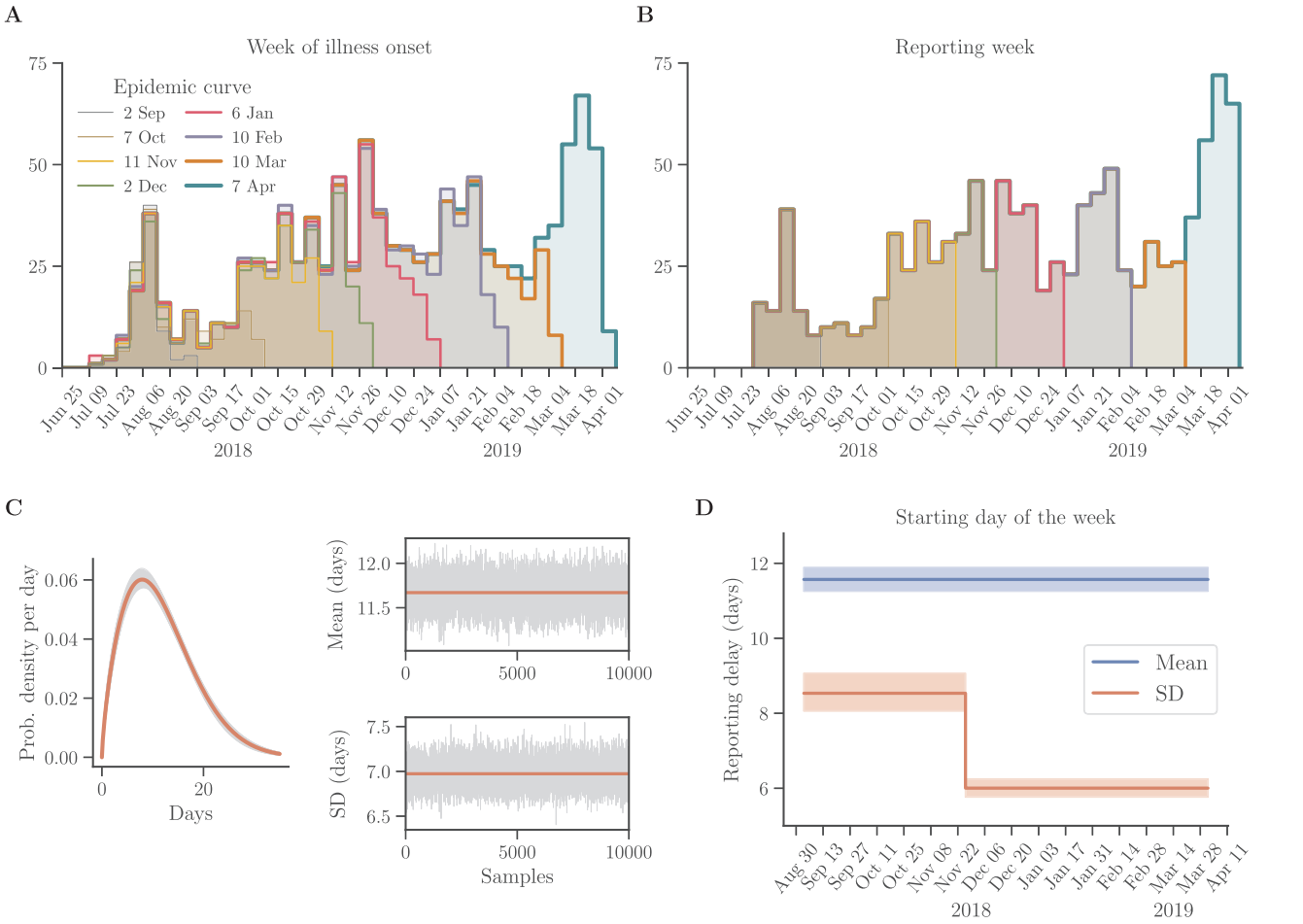

Fig. 1. Epidemic curves of Ebola virus disease in the Democratic Republic of the Congo (DRC) from July 2018 – April 2019 and estimates of the mean and standard deviation (SD) of reporting delay. The epidemic curves for laboratory confirmed cases of Ebola by week of illness onset (A) and by the week of reporting week (B), as reported by the World Health Organization, Regional Office for Africa (WHO AFRO). Among the total 82 available epidemic curves, six epidemic curves including the curves of the first week of each month and the latest curve of the Ebola epidemic in DRC are shown. (C) Mean (11.7 days, 95% CI: 11.6, 11.7) and SD (7.0 days, 95% CI: 6.9, 7.1) of reporting delay via Weibull distribution that does not account for the change point of the statistical moment, based on the latest epidemic curve (i.e., reported on 7 April 2019). Red line indicates the estimates and grey shades present 95% confidence interval. (D) Using the best delay function with change point (i.e., 26 November 2018), stable mean and time-dependent SD, estimated mean was not changed, but SD decreased after the change point from 8.5–6.0 days. Lines and shade present estimates and 95% confidence intervals of each estimate, respectively (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

a negative binomial distribution, as assumed elsewhere (Champredon et al., 2018; Cowling et al., 2010). The distribution has two parameters: the number of already reported cases $c_{t,x}$ and the probability of not reported $p_{t,x}(\theta) = 1 - F(t;\theta)$. As Champredon et al. (2018) indicated in an elegant manner, $c_{t,x}$ describes the number of "failures" and $p_{t,x}(\theta)$ represents the probability of "success" according to the failure-counting definition of the negative binomial distribution. The 95% CI of the total number of cases including reported and not yet reported was computed to quantify the uncertainty in parameters θ sampled from multivariate normal distribution and the total number of cases sampled from negative binomial distribution.

To assess how well the delay distribution allows us to nowcast the total number of cases, we nowcasted recent cases for three epidemic curves as a function of the illness onset time on 11 November, 2 December 2018 and 6 January 2019 and compared them to the latest epidemic curve reported on 7 April 2019. The performance of the nowcasting for the last four data points, i.e. $(x - 4\,\text{weeks}) < t \leq x$ was assessed by the absolute and the relative root-mean-square error (RMSE), the bias, the log-likelihood value, the log scores, following a proper scoring rule from "Forecast the Influenza Season Collaborative Challenge" (FluSight) (Reich et al., 2019). We followed the classical definition:

$$\text{RMSE}^{(\text{abs})}_{x,x'} = \sqrt{\text{mean}(\varepsilon^2_{t,x,x'})}, \tag{4}$$

where $\varepsilon_{t,x,x'} = \text{median}(\hat{c}_{t,x}) - c_{t,x'}$ is the difference between nowcasted counts based on the epidemic curve x and counts reported at the latest epidemic curve x'. The values of the relative RMSE and bias were defined according to Li et al. (Li et al., 2018):

$$\text{RMSE}^{(\text{rel})}_{x,x'} = \sqrt{\text{mean}(\varepsilon^2_{t,x,x'})}, \ \text{bias}_{x,x'} = \text{median}(\varepsilon_{t,x,x'}), \tag{5}$$

respectively, where the log-transformed difference is $\varepsilon_{t,x,x'} = \log\left(\text{median}(\hat{c}_{t,x})/c_{t,x'}\right)$.

We followed Reich et al (2019) to define the "proper score" for which any predicted value $\hat{c}_{t,x}$ with a deviation of no more than 3 cases from the observed number $c_{t,x'}$ is regarded as accurate. Obtaining $f_x(\hat{c}_{t,x}|c_{\circ,x'})$, a predicted density function for the snapshot x, and target $\hat{c}_{t,x}$, conditional on the data $c_{\circ,x'}$, the modified log score in week t is calculated as $\log\int_{c_{t,x'}-3}^{c_{t,x'}+3} f_x(z|c_{\circ,x'})\text{d}z$. The resulting log score for the nowcasted epidemic curve x is defined as the average of all log scores over the last four data points:

$$LS_x = \frac{1}{4}\sum\nolimits_{x-4\,<\,t\leq x} \log\int_{c_{t,x'}-3}^{c_{t,x'}+3} f_x(z|c_{\circ,x'})\,\text{d}z.$$

The proper score is given by the exponent of the log score: $AS_x = e^{LS_x}$. Contrary to LS_x, a scale of AS_x can be explicitly interpreted: it yields a score equivalent to the geometric mean of the probabilities assigned to the observed outcome $x^{'}$ (Reich et al., 2019).

2.4. Forecasting implementation

We project the number of unreported cases $\tilde{c}_{t,x} = \hat{c}_{t,x} - c_{t,x}$ $(t \leq x)$ on two future weeks, i.e., $t_1 = (x + \tau_1)$, and $t_2 = (x + \tau_2)$ where $\tau_1 = 7$ and $\tau_2 = 14$ days. It results in sampling the cases from a multinomial distribution:

$$c_{\{t_k\},x} \tilde{} \sum_{t \leq x} \text{Multinomial}(size = \tilde{c}_{t,x},\ prob = \{p_{t,t_k,x}\}),\ k = \{1,\ 2\}, \tag{6}$$

where the vector of probabilities is constituted by the values:

$$p_{t,t_1,x} = \frac{F(\tau_1 + x - t + 1;\ \theta) - F(x - t + 1;\ \theta)}{1 - F(x - t + 1;\ \theta)},\ p_{t,t_2,x} = \frac{F(\tau_2 + x - t + 1;\ \theta) - F(\tau_1 + x - t + 1;\ \theta)}{1 - F(x - t + 1;\ \theta)}.$$

The obtained (resampled) numbers from a multinomial distribution, i.e., $c_{\{t_k\},x}$ constitute the so-called "forecast" with a lower bound, because they do not involve any new incidence after the reporting time of epidemic curve x. We thus consider several alternative scenarios to predict future in a mechanistic manner. Specifically, we consider predicting the time evolution of the effective reproduction number R_t in week t and also we construct polynomials of different degree to impute future R_t. As for the former, moving average of R_t for two and three past weeks or assuming an identical R_t to the latest estimate were considered. When polynomials were employed, we considered linear, quadratic and cubic functions to model R_t over the past five time points to predict R_t for two future weeks. Lastly, as part of the evaluation metrics of forecasting, we employed the median log-likelihood that is assumed to follow a Poisson distribution:

$$\text{loglk}_{x,x'} = \sum_{t \leq x'} c_{t,x'} \log(\text{median}(\hat{c}_{t,x})) - (\text{median}(\hat{c}_{t,x})) - \log\left(c_{t,x'}!\right). \tag{7}$$

To obtain the value of R_t for any week $t \leq x^{'}$, we assumed that the projected counts of laboratory confirmed cases follow a renewal process:

$$E(\hat{c}_{t,x'}) = R_t (\hat{c} * g)_{t,x'}, \tag{8}$$

where the symbol "*" stands for the convolution operator, i.e.:

$$(\hat{c} * g)_{t,x'} = \sum_{h=1}^{t-1} \hat{c}_{t-h,x'} g_h.$$

where g_h represents the probability mass function of serial interval in week h derived from the cumulative distribution function of the continuous gamma distribution function, $G(h)$, i.e., $g_h = G(h) - G(h-1)$ for $h > 0$ with mean at 15.3 days and standard deviation at 9.3 days (WHO Ebola Response Team, 2014). The estimate of the effective reproduction number can be obtained by maximizing the likelihood that is assumed to follow a Poisson distribution, i.e.,

$$\hat{L}(R_t;\ \hat{c}_{t,x'}) = \prod_{t \leq x'} \frac{(E(\hat{c}_{t,x'}))^{\hat{c}_{t,x}} \exp(-E(\hat{c}_{t,x'}))}{\hat{c}_{t,x'}!}, \tag{9}$$

and, in the analytical form of estimator as

$$R_t = \frac{\sum_{t \leq x'} \hat{c}_{t,x'}}{\sum_{t \leq x'} (\hat{c} * g)_{t,x'}}, \tag{10}$$

where both sums in numerator and denominator are taken only over the set of non-zero convolutions $(\hat{c} * g)_{t,x}$. The estimator of the so-called instantaneous reproduction number in (10) based on renewal equation was employed by numerous researchers elsewhere (White and Pagano, 2008; Nishiura and Chowell, 2009), informing the time-dependent average of the number of secondary cases produced by a single primary case and reflecting the time-dependent impact of public health interventions on the transmission dynamics.

In a similar manner to nowcasting procedures, forecasting was implemented to impute R_t for two future weeks. The new incidence in two future weeks, i.e., $\check{c}_{t_1,x'}$ and $\check{c}_{t_2,x'}$ were calculated by using the renewal equation. By sampling the cases from a multinomial distribution, the incidence, i.e., $\hat{c}_{t_1,x'}$ and $\hat{c}_{t_2,x'}$ were randomly sampled:

$$C_{t_1,x'} \tilde{} \text{Multinomial}(size = \check{c}_{t_1,x'},\ prob = \{F(\tau_1;\ \theta),\ F(\tau_2;\ \theta) - F(\tau_1;\ \theta),\ 1 - F(\tau_2;\ \theta)\}),$$

$$C_{t_2,x'} \tilde{} \text{Multinomial}(size = \check{c}_{t_2,x'},\ prob = \{F(\tau_2 - \tau_1;\ \theta),\ 1 - F(\tau_2 - \tau_1;\ \theta)\}),\ \hat{c}_{t_1,x'} = C^{(1)}_{t_1,x'},\ \hat{c}_{t_2,x'} = C^{(2)}_{t_1,x'} + C^{(1)}_{t_2,x'},$$

where $C^{(j)}_{t_o,x'}$ is the j -th element of the sample $C_{t_1,x'}$.

Each forecasting method was evaluated using several metrics described above, i.e., $\text{RMSE}^{(\text{abs})}$, $\text{RMSE}^{(\text{rel})}$, bias, proper score, and log-likelihood. The performance of each forecast was calculated by comparing the forecasted numbers to the differences in cumulative counts of reported cases between two subsequent reports. Time periods $\tau_1^{'}$ and $\tau_2^{'}$ describe two time intervals between publication of the current and two follow-up reports.

2.5. Data sharing

The dataset and code script are available online at https://github.com/aakhmetz/EbolaForecast2019Epidemics

3. Results

Six snapshots by reporting date, i.e., those reported on 7 October, 11 November, 2 December, 6 January 2019, 17 February and 31 March, and also by week of illness onset are shown in Figs. 1A and B. These curves were used for visual exposition of our nowcasting and forecasting procedures, out of the total of 82 snapshots as of the first week of April 2019. Using all 82 curves, four different distributions of reporting delay (i.e., Weibull, gamma, log-normal and exponential) were independently fitted to the data and the goodness of fit was compared (Table 1). Weibull distribution with the mean of 11.7 days (95% CI: 11.6, 11.7) and standard deviation of 7.0 days (95% CI: 6.9, 7.1) was selected as the best model, yielding the minimum AICc = 9636.8 (Fig. 1C).

Fig. 1D shows the mean and standard deviation (SD) of the reporting delay. Using the latest epidemic curve, four possible combinations of mean and SD were explored and compared (Table S1), allowing the change point to vary over time, and the model with stable mean and time-dependent SD yielded the minimum AICc = 9423.7. Consequently, it was identified that not the mean but the SD has changed

Table 1
Comparison of four candidate distributions to describe the reporting delay of Ebola virus disease in Democratic Republic of the Congo.

Distribution	Mean	Standard deviation	AICc
Weibull	11.7	7.0	9636.8
Gamma	12.4	8.6	9713.6
Lognormal	13.4	12.3	9950.5
Exponential	15.1	_	10372.0

Comparison of AIC with a correction for small sample sizes (AICc) values for four different distributions. Each AICc was calculated as AICc= $2k - 2\ln(L) + 2(k^2 + k)/(n - k - 1)$, where L is the loglikelihood of reporting delay function and k represent the number of parameters and n is the sample size.

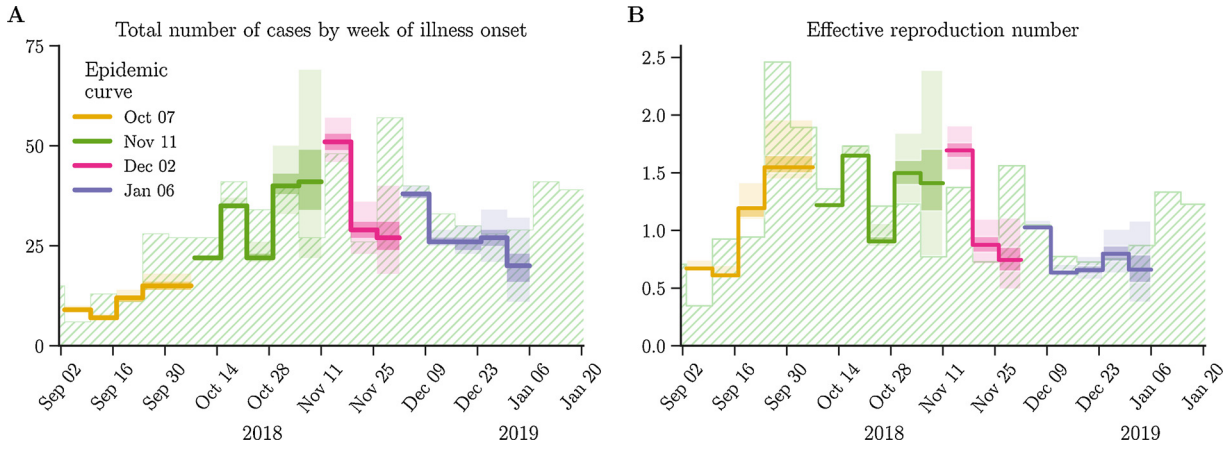

Fig. 2. Nowcasted incidence and the effective reproduction number in the Democratic Republic of the Congo (DRC) from July 2018 – January 2019. (A) Nowcasted incidence data by week of illness onset for four epidemic curves. Each line indicates the median, while dark and light shades show the interquartile range and 95% confidence intervals, respectively. (B) The effective reproduction number of Ebola virus disease outbreak in the DRC as estimated from the nowcasted number of cases. Median, interquartile range and the 95% confidence intervals of the effective reproduction number are shown as in (A). Green striped shades represent values that were reported on 7 April 2019, i.e. the latest epidemic curves (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

from 26 November 2019, shortened from 8.5 (95% CI: 8.1, 9.0) to 6.0 (95% CI: 5.8, 6.3) days. The identified change is consistent with the retrospective traceback of the illness onset time that took place in November, and subsequently, the variation of reporting delay was reduced, which was consistent with the analysis of reporting delay distribution over time (Figure S1).

To visually assess the performance of nowcasting, Fig. 2A shows the comparison between observed and nowcasted data by week of illness onset for four epidemic curves, i.e., those reported on 7 October 2018, 11 November 2018, 2 December, and 6 January 2019. Nowcasted cases involved a large uncertainty using the epidemic curve reported on 11 November, but the cases in the first nowcasted week successfully contained the observed value within its 95% prediction interval. While later nowcasting mostly contained the observed value within the 95% prediction interval, the second nowcasted week using the epidemic curve reported on 2 December was an underestimate, failing to capture an irregular increase in the incidence. The latest nowcasting result was not assessed due to the absence of observed data, but this performance will be seen in near future. The effective reproduction number was also estimated accounting for the reporting delay (Fig. 2B). Table 2 summarizes the performance of the nowcasting according to four different dates of reporting. Evaluation metrics did not reveal monotonous increase or decrease, reflecting several abrupt increases and decreases in the incidence in the observed epidemic curve.

Selecting six epidemic curves distantly separated from each other and among those that were reported on Sundays in the first six month of the outbreak, short-term forecasts were obtained for two future weeks (Figure S2). Comparing evaluation metrics by different model and over time (Table 3 and Figure S3), moving average of R_t yielded better performance of both validity and reliability than others over time (Fig. 3). However, the performance of prediction was differently compared by the week of forecasting. Lower bound always underestimated the future incidence, and the use of interpolated R_t underperformed compared with the use of moving average. However, using moving average of R_t, the second forecasted week yielded a large uncertainty (Fig. 4), and this was even more so using the polynomials. Incorporating the change point into reporting delay distribution yielded improved fit to the observed data for later time snapshots than using the stable (constant) delay distribution (Figures S4 and S5).

4. Discussion

The present study tackled nowcasting and forecasting of the EVD outbreak in DRC in real time. Unlike the large epidemic in West Africa from 2014-16, which even evoked the idea of RAPIDD Ebola forecasting challenge (Viboud et al., 2018), the epidemic size of the latest DRC outbreak remained relatively small, while the shape of the epidemic curve was not smooth and unimodal. In addition, whereas spatial heterogeneity was anticipated, having a constant access to spatially structured snapshot data for multiple time points has been too challenging. Nevertheless, multimodal, small-size, and spatially dependent epidemic curve made it too difficult to apply a homogeneous model for future forecasting purpose, because it may rely on a depletion of susceptible individuals due to natural infections. Thus, we decided to take a semi-structured approach, accounting for the reporting delay which can depend on time. The estimated mean reporting delay from a previous study analyzing the Ebola outbreak in DRC from May 2018 – January 2019 was 3.3 weeks in September 2018 and declined to 1.7 weeks with narrow confidence intervals in January 2019 (Tariq et al., 2019), which shows consistency with that of the present study (i.e.,1.7 weeks). As a result, nowcasting was successfully implemented, mostly containing future observed values within the 95% prediction intervals,

Table 2
Performance of the nowcasting by different reporting date.

Epidemic curve (reporting date)	Change point	$RMSE^{(abs)}$	$RMSE^{(rel)}$	Proper score	Bias	AICc	The latest nowcasted value included the actual value
7 Oct 2018	17 Sep 2018	20.1	0.3	0.0	0.1	57.4	No
11 Nov 2018	5 Nov 2018	9.7	0.1	0.3	0.0	101.5	Yes
12 Dec 2018	5 Nov 2018	17.3	0.2	0.0	0.1	74.0	Yes
6 Jan 2019	5 Nov 2018	5.0	0.1	0.4	0.0	33.5	Yes

Epidemic curve shows the date on which the corresponding incidence was released. The change point indicates the change in the parameter (i.e. standard deviation) of delay distribution. All characteristics $RMSE^{(abs)}$, $RMSE^{(rel)}$, proper score, bias, and AICc are derived only for the last four data points prior the reporting date.

Table 3
Forecasting performance by different reporting date and extrapolated reproduction number for two future weeks.

Epidemic curve (reporting date)	Extrapolation method of the reproduction number	$RMSE^{(abs)}$	$RMSE^{(rel)}$	Proper score	Bias	AICc	Weight	The latest nowcasted value included the actual value
6 Jan 2019	Lower bound: $R_{tk} \equiv 0$	23.6	0.5	0	−0.4	79.9	0.0	Yes/No
	Identical to the last R_t: $R_{tk} = R_x$	15.6	0.2	0.1	−0.2	25.2	0.1	Yes/No
	Moving average of last two R_t: $R_{t_k} = \frac{R_x + R_{x-1}}{2}$	14.2	0.2	0.1	−0.1	20.8	0.7	Yes/No
	Moving average of last three R_t: $R_{t_k} = \frac{R_x + R_{x-1} + R_{x-2}}{3}$	14.9	0.2	0.1	−0.1	22.9	0.2	Yes/No
	Linear approximation	17.0	0.3	0.1	−0.2	30.5	0.0	Yes/Yes
	Quadratic polynomial	20.5	0.4	0.1	−0.3	50.5	0.0	Yes/Yes
	Cubic polynomial	22.0	0.5	0.1	−0.3	62.7	0.0	Yes/Yes
	Model averaging	14.9	0.2	0.1	−0.1	22.9		Yes/No
16 Dec 2018	Lower bound: $R_{tk} \equiv 0$	13.2	0.3	0.0	−0.3	23.5	0.0	No/No
	Identical to the last R_t: $R_{tk} = R_x$	7.6	0.1	0.2	−0.1	9.0	0.3	Yes/Yes
	Moving average of last two R_t: $R_{t_k} = \frac{R_x + R_{x-1}}{2}$	6.0	0.1	0.3	0.1	7.9	0.6	Yes/Yes
	Moving average of last three R_t: $R_{t_k} = \frac{R_x + R_{x-1} + R_{x-2}}{3}$	9.9	0.2	0.1	0.1	12	0.1	Yes/No
	Linear approximation	12.6	0.3	0.0	−0.3	22.3	0.0	No/No
	Quadratic polynomial	12.6	0.3	0.0	−0.3	22.3	0.0	No/No
	Cubic polynomial	12.6	0.3	0.0	−0.3	22.3	0.0	Yes/Yes
	Model averaging	5.7	0.1	0.2	−0.1	6.9		Yes/Yes
25 Nov 2018	Lower bound: $R_{tk} \equiv 0$	24.9	0.4	0.0	−0.3	67.6	0.0	Yes/No
	Identical to the last R_t: $R_{tk} = R_x$	14.9	0.2	0.1	0.1	17.6	0.7	No/Yes
	Moving average of last two R_t: $R_{t_k} = \frac{R_x + R_{x-1}}{2}$	18.7	0.2	0.0	0.2	22.6	0.1	No/Yes
	Moving average of last three R_t: $R_{t_k} = \frac{R_x + R_{x-1} + R_{x-2}}{3}$	17.0	0.2	0.0	0.2	20.1	0.2	No/Yes
	Linear approximation	24.3	0.4	0.0	−0.3	61.9	0.0	Yes/No
	Quadratic polynomial	24.9	0.4	0.0	−0.3	67.6	0.0	Yes/No
	Cubic polynomial	24.9	0.4	0.0	−0.3	67.6	0.0	Yes/No
	Model averaging	14.8	0.2	0.1	0.1	17.1		No/Yes
28 Oct 2018	Lower bound: $R_{tk} \equiv 0$	16.6	0.4	0.0	−0.3	40.8	0.0	Yes/No
	Identical to the last R_t: $R_{tk} = R_x$	2.5	0.0	0.2	0.0	5.5	0.6	Yes/Yes
	Moving average of last two R_t: $R_{t_k} = \frac{R_x + R_{x-1}}{2}$	9.2	0.1	0.2	0.1	9.3	0.1	Yes/Yes
	Moving average of last three R_t: $R_{t_k} = \frac{R_x + R_{x-1} + R_{x-2}}{3}$	5.7	0.1	0.3	0.0	6.9	0.3	Yes/Yes
	Linear approximation	12.1	0.2	0.1	−0.2	19.7	0.0	Yes/Yes
	Quadratic polynomial	16.6	0.4	0.1	−0.3	40.8	0.0	Yes/Yes
	Cubic polynomial	16.3	0.4	0.0	−0.3	40	0.0	Yes/No
	Model averaging	16.6	0.4	0.0	−0.3	40.8		Yes/No
7 Oct 2018	Lower bound: $R_{tk} \equiv 0$	5.7	0.1	0.2	−0.1	8.5	1.0	Yes/Yes
	Identical to the last R_t: $R_{tk} = R_x$	81.1	0.5	0.0	0.4	152.6	0.0	Yes/No
	Moving average of last two R_t: $R_{t_k} = \frac{R_x + R_{x-1}}{2}$	48.4	0.4	0.0	0.3	78.2	0.0	Yes/No
	Moving average of last three R_t: $R_{t_k} = \frac{R_x + R_{x-1} + R_{x-2}}{3}$	35.6	0.3	0.1	0.3	52.0	0.0	Yes/No
	Linear approximation	152.0	0.7	0.0	0.6	332.4	0.0	Yes/No
	Quadratic polynomial	228.0	0.8	0.0	0.7	536.4	0.0	Yes/No
	Cubic polynomial	311.9	0.9	0.0	0.8	767.3	0.0	Yes/No
	Model averaging	39.2	0.4	0.1	0.3	59.1		Yes/No
16 Sep 2018	Lower bound: $R_{tk} \equiv 0$	7.3	0.8	0.1	−0.7	36.5	0.0	Yes/No
	Identical to the last R_t: $R_{tk} = R_x$	5.1	0.4	0.3	−0.3	12.2	0.0	Yes/Yes
	Moving average of last two R_t: $R_{t_k} = \frac{R_x + R_{x-1}}{2}$	4.1	0.3	0.4	−0.3	8.1	0.3	Yes/Yes
	Moving average of last three R_t: $R_{t_k} = \frac{R_x + R_{x-1} + R_{x-2}}{3}$	3.5	0.2	0.4	−0.2	6.5	0.7	Yes/Yes
	Linear approximation	6.7	0.6	0.2	−0.6	24.6	0.0	Yes/Yes
	Quadratic polynomial	6.7	0.6	0.2	−0.6	24.6	0.0	Yes/Yes
	Cubic polynomial	6.3	0.5	0.2	−0.5	22		Yes/Yes

Epidemic curve shows the date on which the corresponding incidence was released. Weight for each epidemic curve i was calculated as $\exp(-0.5\Delta_i))/\sum_i \exp(-0.5\Delta_i)$ where $\Delta_i = AIC_C^i - \min(AIC_c)$ and AIC_C^i represents AIC_C for epidemic curve i and $\min(AIC_c)$ is the minimum value of AIC_C among all epidemic curves. All characteristics $RMSE^{(abs)}$, $RMSE^{(rel)}$, proper score, bias, AICc and weight are derived for two future data points. Included the latest values in the 95% prediction interval shows if the observed data in first and second forecasted weeks were included within the forecasted 95% prediction intervals. Note that the forecast on 16 September 2018 is not accompanied by model averaging results, because this time point does not have AICc of prior forecast that acts as the weight of different models.

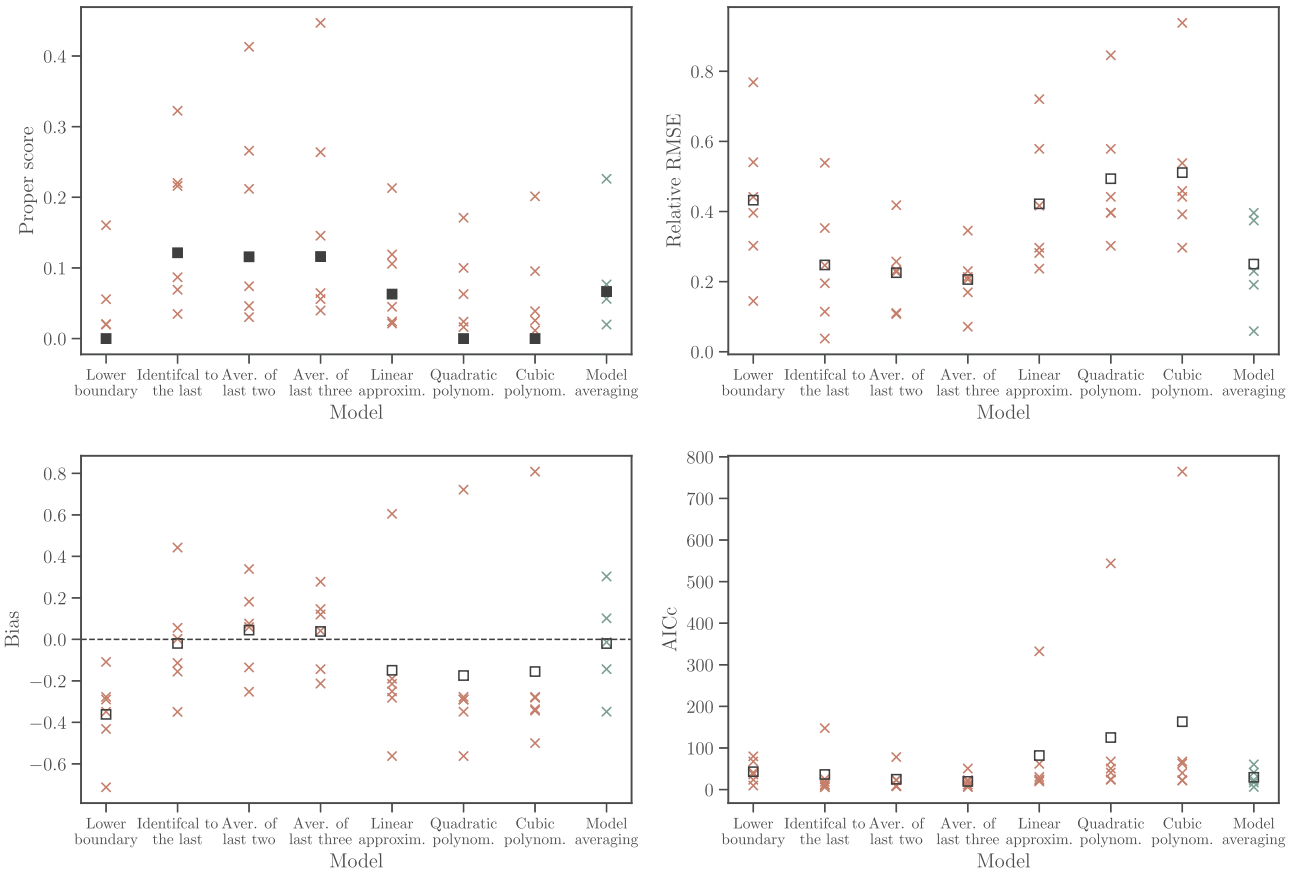

Fig. 3. Comparative performance of forecasts using different methods of future extrapolations. Horizontal axis from left to right: Forecasted number of confirmed cases, considering only the lower bound, i.e. assuming that R_t is zero; forecasted number of confirmed cases using an identical R_t to the latest estimate, using moving average over two latest estimates of R_t, over three latest estimates of R_t ; using a linear approximation of R_t, quadratic polynomial of R_t, and cubic polynomial in forecasting R_t for two future weeks, and lastly, using averaging of all mentioned methods where the weight to each methods are assigned according to Akaike Information Criterion (AIC) values of the preceding forecast. Each crossed point corresponds to a particular epidemic curve (in total six epidemic curves that were subject for forecasting exercise). Black squares stand for geometric (filled) and arithmetic (unfilled) means over all forecasts.

but there were failures when the incidence abruptly changed over time. Forecasting exercise was also performed, and similar to nowcasting results, there were occasional failures, but high average forecast scores demonstrated the validity and reliability through our extrapolation approach, using the moving average of the reproduction numbers for the latest few weeks.

We must recognize that the latest outbreak in DRC from 2018-19 involved not only small epidemic size with spatial heterogeneities in a data-limited setting but also a large-scale vaccination campaign in real time. Without knowing the details of vaccination in real time, we have lost the opportunity to explicitly offer forecasts using a fully mechanistic model. Instead, we took a simplistic approach to extrapolating a part of the mechanistic model, i.e., R_t in the future, using several candidate functions. That is, our nowcasting procedure purely rested on a statistical model, handling the reporting delay distribution, but forecasting has involved the extrapolation of future R_t. Even without extrapolation, our nowcasting exercise has shown that its success depends on the presence of abrupt change in the incidence. Qualitatively, similar findings were also the case for forecasting, successfully obtaining valid and reliable forecasts of two future weeks. Among various possibilities to extrapolate future R_t, the forecast using a moving average of R_t for the past few weeks outperformed others, including that assumes $R_t = 0$ or an identical R_t to the latest estimate (Fig. 4).

The present study was originally motivated by our desire to incorporate the proposed tool of nowcasting and short-term forecasting into the surveillance practice. Surveillance and mathematical modelling are two complementary instruments in the toolbox of infectious disease epidemiologists, and we trust that adding the proposed nowcasting and forecasting framework will truly improve the situation awareness. In this regard, we have demonstrated that (i) devising the proposed model, the situation awareness of the latest two weeks can be considerably improved by accounting for the reporting delay (Tariq et al., 2019) and (ii) even with this size of the outbreak, short-term forecasting could be tackled by imposing small extrapolations to R_t. Only by adding these two pieces of information to the surveillance data, we believe that the information that we can gain from the real time update of "situation report" can be considerably improved.

While the present study addressed reporting delay distribution, it must be noted that the present study was not able to adjust the reporting rate (i.e. under-ascertainment). Unfortunately, we suspect that many cases remain still unreported because of socially and culturally difficult nature of the outbreak, e.g. a high fraction of death cases in community has remained non-confirmed (Shuchman, 2019). In this regard, it must be remembered that if the reporting rate changed over time, our modeling results could be considerably biased, which can result in both over- and under-estimation of R_t.

There are four other limitations to be remembered. First, we did not include probable cases in estimating R_t which could have resulted in underestimation especially during the early epidemic period. Second, it is well known that spatial effects of spread are not fully captured by non-spatial models (Keeling and Rohani, 2011), but we tackled the issue of nowcasting and forecasting as if the data were non-spatial. Third, we were not able to track susceptible individuals' density in North Kivu, especially vaccinated individuals, and thus, such

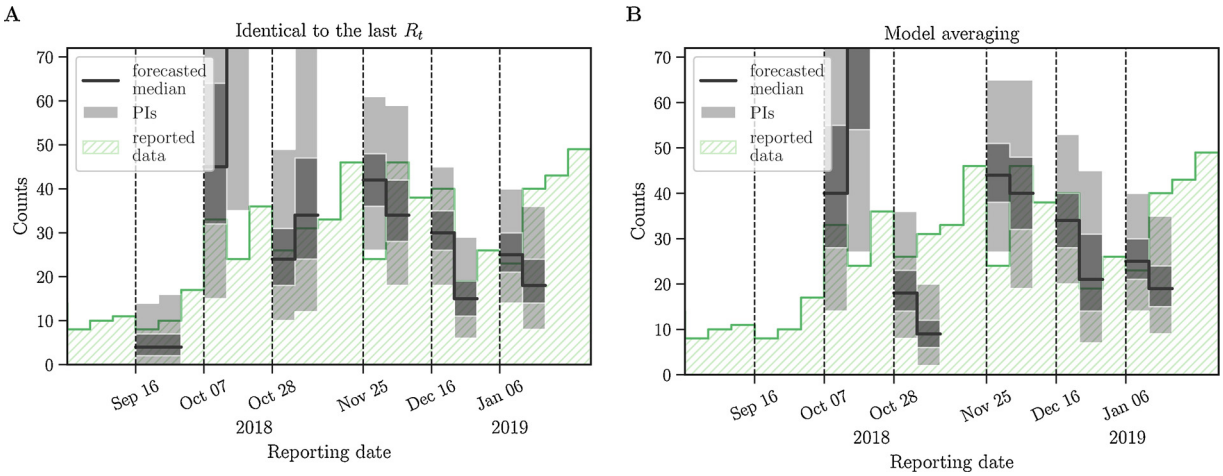

Fig. 4. Short-term forecast of the incidence using the best performed model with highest geometric mean of the proper score (A) and using the model averaged over all methods (B). Forecasted incidence by week of reporting for four epidemic curves with the future reproduction number identical to the last value of R_t. (A) Forecasted incidence of averaged over all considered methods with the weights for each method gained from a preceding forecast (B) are shown for two future weeks. In both panels, lines indicate the forecasted median and dark and light shades show the interquartile range and the 95% prediction intervals (PIs), respectively. Green striped shades represent the reported laboratory confirmed cases by week of reporting (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

information was not incorporated into the model building process in a direct manner. Fourth, depending on the locality, the outbreak from 2018-19 is featured by political instability, community resistance (e.g. vaccine refusals at locality) and conflict, but that point should ideally be explored in relation to the reporting rate and reporting delay in the future.

Adhering to the surveillance system, reporting delay and reporting coverage make it difficult to capture the epidemic dynamics in a timely manner. Integrating nowcasting and forecasting tool to the ongoing surveillance, we trust that the proposed simplistic model would permit a far better interpretation of the awareness toward the current and future situations of EVD outbreak in the DRC than solely relying on observed data alone. As we have shown, having epidemic curves by both illness onset time and reporting time would allow an explicit estimation of the reporting delay distribution and nowcasting, and thus, real time updates on both dimensions should ideally be considered in updating epidemiological reports.

Acknowledgments

HN received funding support from the Japan Agency for Medical Research and Development (AMED, grant number JP18fk0108050), Japan Society for the Promotion of Science (JSPS) KAKENHI (grant Numbers 16KT0130, 16K15356, 17H04701, 17H05808, and 18H04895), Inamori Foundation, and the Japan Science and Technology Agency (JST) CREST program (JPMJCR1413). HL received funding from the Japan Society for the Promotion of Science (JSPS) KAKENHI (grant number 19K19343). SMJ acknowledges the Japanese government (Monbusho) scholarship support from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan, and BY acknowledges the China Scholarship Council. The funders played no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. ARA is thankful to Jonathan Dushoff (McMaster University, Canada) for helpful discussions.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.epidem.2019.05.002.

References

Barry, A., Ahuka-Mundeke, S., Ali Ahmed, Y., Allarangar, Y., Anoko, J., Archer, B.N., Aruna Abedi, A., Bagaria, J., Belizaire, M.R.D., Bhatia, S., Bokenge, T., Bruni, E., Cori, A., Dabire, E., Diallo, A.M., Diallo, B., Donnelly, C.A., Dorigatti, I., Dorji, T.C., Escobar Corado Waeber, A.R., Fall, I.S., Ferguson, N.M., FitzJohn, R.G., Folefack Tengomo, G.L., Formenty, P.B.H., Forna, A., Fortin, A., Garske, T., Gaythorpe, K.A., Gurry, C., Hamblion, E., Harouna Djingarey, M., Haskew, C., Hugonnet, S.A.L., Imai, N., Impouma, B., Kabongo, G., Kalenga, O.I., Kibangou, E., Lee, T.M.-H., Lukoya, C.O., Ly, O., Makiala-Mandanda, S., Mamba, A., Mbala-Kingebeni, P., Mboussou, F.F.R., Mlanda, T., Mondonge Makuma, V., Morgan, O., Mujinga Mulumba, A., Mukadi Kakoni, P., Mukadi-Bamuleka, D., Muyembe, J.-J., Bathé, N.T., Ndumbi Ngamala, P., Ngom, R., Ngoy, G., Nouvellet, P., Nsio, J., Ousman, K.B., Peron, E., Polonsky, J.A., Ryan, M.J., Touré, A., Towner, R., Tshapenda, G., Van De Weerdt, R., Van Kerkhove, M., Wendland, A., Yao, N.K.M., Yoti, Z., Yuma, E., Kalambayi Kabamba, G., Lukwesa Mwati, J., de, D., Mbuy, G., Lubula, L., Mutombo, A., Mavila, O., Lay, Y., Kitenge, E., 2018. Outbreak of Ebola virus disease in the Democratic Republic of the Congo, April–may, 2018: an epidemiological study. Lancet 392, 213–221. https://doi.org/10.1016/S0140-6736(18)31387-4.

Champredon, D., Li, M., Bolker, B.M., Dushoff, J., 2018. Two approaches to forecast Ebola synthetic epidemics. Epidemics 22, 36–42. https://doi.org/10.1016/j.epidem.2017.02.011.

Cowling, B.J., Lau, M.S.Y., Ho, L.-M., Chuang, S.-K., Tsang, T., Liu, S.-H., Leung, P.-Y., Lo, S.-V., Lau, E.H.Y., 2010. The effective reproduction number of pandemic influenza: prospective estimation. Epidemiology 21, 842–846. https://doi.org/10.1097/EDE.0b013e3181f20977.

Finger, F., Funk, S., White, K., Siddiqui, R., Edmunds, W.J., Kucharski, A.J., 2018. Real-time analysis of the diphtheria outbreak in forcibly displaced Myanmar nationals in Bangladesh. bioRxiv. https://doi.org/10.1101/388645.

Funk, S., Camacho, A., Kucharski, A.J., Eggo, R.M., Edmunds, W.J., 2018. Real-time forecasting of infectious disease dynamics with a stochastic semi-mechanistic model. Epidemics 22, 56–61. https://doi.org/10.1016/j.epidem.2016.11.003.

Hahné, S., Donker, T., Meijer, A., Timen, A., van Steenbergen, J., Osterhaus, A.D., van der Sande, M., Koopmans, M., Wallinga, J., Coutinho, R., the Dutch New Influenza A(HN)v, Inve, C., 2009. Epidemiology and control of influenza A(H1N1)v in the Netherlands: the first 115 cases. Eurosurveillance 14. https://doi.org/10.2807/ese.14.27.19267-en.

Keeling, M.J., Rohani, P., 2011. Modeling Infectious Diseases in Humans and Animals. Princeton University Press, New Jersey.

Li, M., Dushoff, J., Bolker, B.M., 2018. Fitting mechanistic epidemic models to data: a comparison of simple Markov chain Monte Carlo approaches. Stat. Methods Med. Res. 27, 1956–1967. https://doi.org/10.1177/096.

Nishiura, H., Chowell, G., 2009. The effective reproduction number as a prelude to statistical estimation of time-dependent epidemic trends. In: Chowell, G., Hyman, J.M., Bettencourt, L.M.A., Castillo-Chavez, C. (Eds.), Mathematical and Statistical Estimation Approaches in Epidemiology. Springer, The Netherlands, pp. 103–121.

Ponce, L., Kinoshita, R., Nishiura, H., 2019. Exploring human-animal interface of Ebola virus disease outbreaks. Math. Biosci. Eng. 16 (4), 3130–31432280217747054.

Reich, N.G., Brooks, L.C., Fox, S.J., Kandula, S., McGowan, C.J., Moore, E., Osthus, D., Ray, E.L., Tushar, A., Yamana, T.K., Biggerstaff, M., Johansson, M.A., Rosenfeld, R., Shaman, J., 2019. A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. Proceedings of the National Academy of Sciences of the U.S.A 116 (8), 3146–3154. https://doi.org/10.1016/10.1073/pnas.1812594116.

Reijn, E., Swaan, C.M., Kretzschmar, M.E., van Steenbergen, J.E., 2011. Analysis of timeliness of infectious disease reporting in the Netherlands. BMC Public Health 11. https://doi.org/10.1186/1471-2458-11-409.

Report of a WHO/International Study Team, 1978. Ebola haemorrhagic fever in Sudan, 1976. Bull. World Health Organ. 56 (2), 247.

Shuchman, M., 2019. Logistical challenges in the DR Congo Ebola virus response. Lancet 393, 117–118. https://doi.org/10.1016/S0140-6736(19)30076-5.

Tariq, A., Roosa, K., Mizumoto, K., Chowell, G., 2019. Assessing reporting delays and the effective reproduction number: the 2018-19 Ebola epidemic in DRC, May 2018-January 2019. Epidemics. https://doi.org/10.1016/j.epidem.2019.01.003.

Tsuzuki, S., Lee, H., Miura, F., Chan, Y.H., Jung, S., Akhmetzhanov, A.R., Nishiura, H., 2017. Dynamics of the pneumonic plague epidemic in Madagascar. August to October 2017. Eurosurveillance 22. https://doi.org/10.2807/1560-7917.ES.2017.22.46.17-00710.

Viboud, C., Sun, K., Gaffey, R., Ajelli, M., Fumanelli, L., Merler, S., Zhang, Q., Chowell, G., Simonsen, L., Vespignani, A., 2018. The RAPIDD ebola forecasting challenge: synthesis and lessons learnt. Epidemics 22, 13–21. https://doi.org/10.1016/j.epidem.2017.08.002.

White, L.F., Pagano, M., 2008. A likelihood-based method for real-time estimation of the serial interval and reproductive number of an epidemic. Stat. Med. 27 (16), 2999–3016.

White, L.F., Wallinga, J., Finelli, L., Reed, C., Riley, S., Lipsitch, M., Pagano, M., 2009. Estimation of the reproductive number and the serial interval in early phase of the 2009 influenza A/H1N1 pandemic in the USA. Influenza Other Respir. Viruses 3, 267–276. https://doi.org/10.1111/j.1750-2659.2009.00106.x.

WHO, 2018a. Ebola Virus Disease, Democratic Republic of the Congo, External Situation Report 17. World Health Organization.

WHO, 2018b. Ebola Virus Disease, Democratic Republic of the Congo, External Situation Report 01. World Health Organization.

WHO, 2018c. Ebola Virus Disease, Democratic Republic of the Congo, External Situation Report 09. World Health Organization.

WHO, 2018d. Ebola Virus Disease, Democratic Republic of the Congo, External Situation Report 16. World Health Organization.

WHO, 2018e. Ebola Virus Disease, Democratic Republic of the Congo, External Situation Report 22. World Health Organization.

WHO, 2018f. Ebola Virus Disease, Democratic Republic of the Congo, External Situation Report 15. World Health Organization.

WHO, 2019. WHO AFRO Situation Reports and Disease Outbreak News. WHO AFRO (Accessed, 14 February 2019). https://www.afro.who.int/health-topics/ebola-virus-disease/.

WHO AFRO, 2019. Weekly Bulletins on Outbreaks and Other Emergencies. WHO AFRO (Accessed 14, February 2019). https://www.afro.who.int/health-topics/disease-outbreaks/outbreaks-and-other-emergencies-updates/.

WHO Ebola Response Team, 2014. Ebola virus disease in West Africa—the first 9 months of the epidemic and forward projections. N. Engl. J. Med. 371 (16), 1481–1495. https://doi.org/10.1056/NEJMc1413884.